Relatório Parcial – Iniciação Científica

Estudo e Aplicação de Modelos de Previsão no Contexto de Séries de Vazões de Rios

Submetido à
Pró-Reitoria de Pesquisa da Unicamp

Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
CEP 13083-852, Campinas, São Paulo (SP)

Candidato: Daniel da Costa Nunes Resende Neto
Orientador: Prof. Levy Boccato

1 Introdução

Uma série temporal é uma sequência de medidas feitas ao longo do tempo sobre um fenômeno de interesse. Em várias áreas do conhecimento, tais como engenharia, economia, matemática aplicada e computação, diversas séries temporais são particularmente relevantes não só pela perspectiva de análise e interpretação de seus históricos passados, mas também pela possibilidade de estimar seus valores futuros. Este último desafio dá origem ao problema de predição de séries temporais [1].

Um cenário em que o problema de predição de séries temporais aparece com destacada importância está associado a medidas de vazões de rios, a fim de estabelecer um planejamento energético mais eficiente e seguro para uma determinada região.

Esta primeira parte do trabalho teve como objetivo estudar e aplicar dois modelos de previsão clássicos e lineares: o auto-regressivo (AR) e o auto-regressivo de médias móveis (ARMA), sobre a série de vazões de Água Vermelha (do dia 01 de janeiro de 2000 ao dia 31 de dezembro de 2015). Os modelos estão detalhados nas seções 2.1 e 2.3 [1].

Um ponto que exigiu um estudo adicional está relacionado à implementação do modelo ARMA. Uma vez que não há solução fechada para os coeficientes do modelo, no sentido de mínimo erro quadrático médio (EQM), foi necessário lançar mão de um algoritmo de otimização inspirado no processo de evolução natural [2], combinado com uma técnica de nicho ecológico [3, 4].

Por fim, foi criado um quadro comparativo entre ambas as técnicas estudadas. Os resultados obtidos são brevemente apresentados e discutidos na Seção 4, concluindo, assim, esta primeira etapa da pesquisa.

2 Modelos de Previsão

2.1 AR

O modelo mais clássico de previsão é o auto-regressivo (AR, do inglês *auto-regressive*) [1, 5, 6]. No AR , o valor da série no instante n, aqui denotado por $x(n)$, é determinado a partir de uma combinação linear dos valores passados até um instante $n-M-L+1$, onde M determina a ordem do modelo e L o passo de previsão (quantos instantes de tempo à frente pretende-se estimar). Em termos matemáticos, a regra de evolução temporal do modelo AR é dada por:

$$x(n) = a_1 x(n-L) + \ldots + a_M x(n-M-L+1) + \eta(n), \tag{1}$$

onde $a_i, i = 1, \ldots, M$ são os coeficientes do modelo AR que ponderam as amostras passadas da série e $\eta(n)$ denota o erro instantâneo, cuja média é nula e cuja variância (σ_η^2) é constante. Este último termo constitui um ruído branco (em inglês, *white noise*) e também é chamado na literatura de "choque aleatório" [1].

A partir de uma sequência de manipulações [5] e considerando $L = 1$ para simplificar, é possível transformar a equação (1) na forma matricial:

$$\underbrace{\begin{bmatrix} r(0) & r(1) & \ldots & r(M-1) \\ r^*(1) & r(0) & \ldots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r^*(M-1) & r^*(M-2) & \ldots & r(0) \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} r^*(1) \\ r^*(2) \\ \vdots \\ r^*(M) \end{bmatrix}}_{\mathbf{r}}, \tag{2}$$

onde $r(M)$ é o operador de autocorrelação do processo $x(n)$ para um atraso M e $w_k = -a_k, k = 1, \ldots, M$. O sistema em (2) define as chamadas equações de Yule-Walker, cuja solução (fechada) para o vetor de coeficientes $\mathbf{w}$ é dada por [1, 5]:

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{r}. \tag{3}$$

2.2 MA

Outro modelo linear clássico na literatura de séries temporais assume que os valores de $x(n)$ resultam de uma combinação linear dos erros passados, ou, analogamente, de valores passados do próprio ruído branco, como mostra a expressão a seguir:

$$x(n) = \sum_{k=0}^{K} b_k \eta(n - k - J + 1), \tag{4}$$

onde $b_0 = 1$, J é o passo do ruído branco e K representa a ordem do modelo, o qual é conhecido como médias móveis (MA, do inglês *moving-average*) [5, 1].

Diferentemente do AR, não é mais possível determinar uma solução em forma fechada para o modelo MA, de maneira que métodos iterativos e/ou heurísticos de busca devem ser empregados.

2.3 ARMA

A combinação das ideias presentes nos modelos AR e MA dá origem a um modelo linear mais geral, denominado ARMA (do inglês, *auto-regressive moving-average*) [1, 5, 6]. Neste caso, o valor da série temporal no instante n depende linearmente de valores passados da própria série e também do ruído branco, como mostra a expressão abaixo:

$$x(n) = \sum_{k=0}^{M} a_k x(n-k-L+1) + \sum_{k=0}^{K} b_k \eta(n-k-J+1), \tag{5}$$

onde as definições dadas para o modelo AR e MA ainda valem. É pertinente mencionar que o valor de L é comumente igual ao de J.

Devido ao uso implícito do MA no modelo ARMA, este último modelo herda a necessidade do uso de métodos iterativos e/ou heurísticos de busca para encontrar os valores ótimos dos coeficientes da regra linear.

3 Metodologia

A mesma metodologia foi empregada durante o projeto e avaliação dos modelos AR e ARMA, de modo a tornar verossímil a comparação de seus respectivos desempenhos. Por conseguinte, a implementação de ambos modelos foi separada em duas etapas: a etapa de validação e a de teste.

Na etapa de validação, os primeiros 14 anos foram divididos em duas partes: os dados de treinamento (01 de janeiro de 2000 até 31 de dezembro de 2009, totalizando 3653 dias) e os dados de validação (01 de janeiro de 2010 até 31 de dezembro de 2013, totalizando 1641 dias). Já na etapa de testes, foram utilizados 2 anos de informação, de 01 de janeiro de 2014 até 31 de dezembro de 2015, totalizando 730 dias.

Nas seguintes seções, cada uma das etapas será detalhada no contexto de cada modelo. Por fim, os métodos serão comparados na Seção 4.

3.1 AR

Para a etapa de validação do modelo AR, desejou-se primeiramente encontrar o valor de atraso M dentro do intervalo I_M que leve ao menor EQM sobre os dados de validação. O intervalo $I_M = [1, 30]$ foi definido a partir do gráfico da autocorrelação parcial da série de vazões.

Para cada valor de M, resolveu-se um sistema descrito na forma matricial por (3) sobre os dados de treinamento. O vetor de parâmetros encontrado $\mathbf{w}$ foi utilizado, então, para prever as vazões diárias dos dados de validação, como mostra a Equação (1). Consequentemente, tem-se o EQM de validação associado àquele atraso.

O processo, então, foi repetido para os outros valores de M em I_M, e o atraso com menor EQM de validação (no caso $M = 29$) foi escolhido para o próximo passo.

Neste, utilizou-se este número de atrasos no preditor AR em um novo treinamento, considerando agora os dados de validação e treinamento concatenados. O conjunto de parâmetros encontrado demarca o fim da etapa de validação.

Na etapa de teste, este último conjunto de parâmetros foi usado para prever todas as saídas (de acordo com a Equação (1)) do conjunto de teste, e, por conseguinte, os valores do EQM e do EAM (erro absoluto médio) finais: $1088{,}6\,(\mathrm{m}^3/\mathrm{s})^2$ e $19{,}8\,\mathrm{m}^3/\mathrm{s}$, respectivamente.

3.2 ARMA

Analogamente, para a etapa de validação do modelo ARMA, desejou-se encontrar primeiramente os valores ótimos para M e K entre as possíveis combinações (explicitadas abaixo). O intervalo I_M foi mantido, enquanto o intervalo $I_K = [1, 60]$ foi escolhido baseado na função de autocorrelação completa da série de vazões.

Os valores de M e K aqui considerados foram: $\mathbf{M} = [1, 5, 15, 30]$ e $\mathbf{K} = [1, 5, 10, 20, 40]$.

Testou-se cada uma das 20 combinações de K e M possíveis. Dada uma combinação de atrasos, criou-se inicialmente uma população de 100 indivíduos (vetores), cujos elementos são gerados a partir de uma distribuição Gaussiana padrão (média nula e variância unitária). Cada indivíduo é representado por um vetor linha de tamanho $G = M + K + 1$ (sendo o último termo o coeficiente independente).

Em seguida, a população entrou em um laço no qual foi submetida às operações típicas de um algoritmo genético básico [3, 4] – a saber, recombinação, mutação e seleção –, até que se completassem 150 repetições (gerações). Essa sequência de operações é resumida a seguir:

I. **variação genética**: a população é submetida a possíveis mutações genéticas (geralmente não drásticas). Em termos computacionais, percorre-se cada um dos indivíduos, e caso ocorra a mutação (com probabilidade p_s), este sofre uma perturbação aleatória em parte de seu conteúdo genético (elementos aleatoriamente selecionados). A perturbação utilizada aqui consiste em somar variáveis aleatórias gaussianas de média nula e variância constante igual a σ_m^2 a cada elemento selecionado. Ela recebe o nome de *mutação gaussiana* [3, 4];

II. **reprodução com herança**: cada indivíduo tem uma probabilidade (p_c) de gerar descendentes que herdam características de seus pais. Caso selecionado, ele e outro indivíduo [preferencialmente do mesmo nicho (ver próximo item); caso não exista, é escolhido da população aleatoriamente] geram juntos dois descendentes $\mathbf{x'}_i$ e $\mathbf{y'}_i$,

cujos parâmetros são combinações convexas dos dois genitores $\mathbf{x}_i$ e $\mathbf{y}_i$. Este operador é conhecido como crossover aritmético [4];

III. **seleção natural**: os indivíduos que avançarão para a próxima geração são selecionados de forma probabilística, com uma chance proporcional às suas respectivas medidas de aptidão (*fitness*). Neste caso, o *fitness* associado a cada indivíduo é o inverso do EQM cometido pelo mesmo ao realizar previsões sucessivas sobre os dados de treinamento. Antes de selecionar os indivíduos, porém, com o objetivo de tentar impedir a convergência de todos eles a uma única região no espaço de busca, introduz-se o conceito de nichos ecológicos (em computação, *niching*). Considera-se que dois indivíduos pertencem a um mesmo nicho se eles ocupam uma mesma região do espaço de busca, isto é, se a distância (Euclidiana) entre eles é inferior a um determinado limiar σ. Neste trabalho, também exploramos a técnica de *clearing*, que atribui todos os recursos de um nicho ao(s) melhor(es) indivíduo(s), enquanto os demais indivíduos passam a ter *fitness* igual a zero. Por fim, a seleção propriamente dita é feita com o auxílio do método *SUS* (do inglês, *Stochastic Universal Sampling*) [3]. Neste operador, quanto maior o *fitness* do indivíduo, maior sua chance de gerar mais descendentes (cópias do próprio indivíduo). Além disso, indivíduos com *fitness* zerados são incapazes de gerar descendentes.

Ao final de cada geração, foram calculados os *fitness* (agora de validação) para cada indivíduo da nova população. O indivíduo de maior *fitness* entre estes foi somente substituído nas próximas gerações caso tivesse surgido nas novas populações outro mais apto. Ao final, obteve-se o menor EQM (o maior *fitness*) de validação para dada combinação de M e K.

Repetiu-se o processo para as outras combinações de M e K. A combinação com menor EQM de validação ao final das gerações foi utilizada para retreinar o modelo ARMA, agora com os dados de treinamento e validação concatenados. O conjunto de parâmetros ótimo obtido demarca o fim da etapa de validação. Por fim, foram obtidas as saídas do modelo ARMA (de acordo com a Equação (5)) para os dados de teste. A partir delas foram calculados os valores de EQM e EAM finais, que foram iguais a $1056{,}3\,(\mathrm{m}^3/\mathrm{s})^2$ e $19{,}7\,\mathrm{m}^3/\mathrm{s}$ respectivamente.

4 Resultados

Como os erros cometidos por ambos os métodos foram bastante semelhantes, apresentamos na Figura 1 somente o gráfico que sobrepõe as vazões originais de Água Vermelha com

aquelas previstas na etapa de teste pelo modelo ARMA (considerando um período de 1 mês).

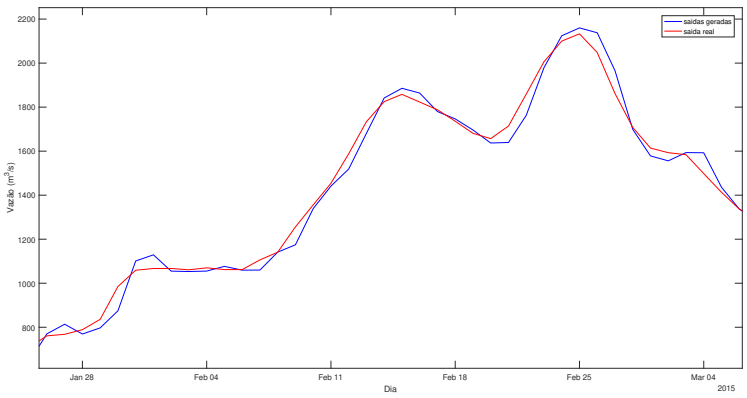

Figura 1: Vazões previstas e vazões reais com *zoom* - ARMA

Pode-se notar que as curvas de vazões estão levemente deslocadas. Ainda assim, os métodos AR e ARMA são altamente capazes de acompanhar a curva original de vazões, implicando que ambos os desempenhos são satisfatórios.

Referências

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley, 2015.

[2] H. V. Siqueira, "Previsão de séries de vazões com redes neurais artificiais e modelos lineares ajustados por algoritmos bio-inspirados," Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, 2009.

[3] L. Boccato, "Aplicação de computação natural ao problema de estimação de direção de chegada," Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, 2010.

[4] T. Bäck, D. Fogel, and Z. Michalewicz, *Evolutionary Computation 1: Basic Algorithms and Operators*. Bristol, UK: Institute of Physics Publishing, 2000.

[5] S. Haykin, *Adaptive filter theory*, 5th ed. Prentice Hall, 2013.

[6] T. M. Bartlett, "Modelagem de séries temporais não-estacionárias através de um modelo arma multimomental baseado em misturas de componentes normais," Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, 2018.